

An asymptotically optimal policy and state-space collapse for the multi-class shared queue

Mark Shifrin

Department of Electrical Engineering
Technion–Israel Institute of Technology

Abstract

We consider a multi-class G/G/1 queue with a finite shared buffer. There is task admission and server scheduling control which aims to minimize the cost which consists of holding and rejection components. We construct a policy that is asymptotically optimal in the heavy traffic limit. The policy stems from solution to Harrison-Taksar (HT) free boundary problem and is expressed by a single free boundary point. We show that the HT problem solution translated into the queue length processes follows a specific *triangular* form. This form implies the queue length control policy which is different from the known $c\mu$ priority rule and has a novel structure.

We exemplify that the probabilistic methods we exploit can be successfully applied to solving scheduling and admission problems in cloud computing.

Keywords: Multiclass G/G/1 queue; Brownian control problems; Harrison-Taksar free boundary problem; Shared buffer system; State dependent priorities

1 Introduction

We consider the problem of finding asymptotically optimal (AO) controls for the multiclass G/G/1 queue with a single *shared* buffer, in heavy traffic. The system is characterized by I classes of arriving tasks. Each task class has its designated queue. The queues share a single buffer, such that the total number of tasks in all queues is limited by its size. We assume that tasks of all classes occupy equally sized storage slots. Upon arrival of a task of class i (with $i \in \{1, \dots, I\}$), a decision maker (DM) may either accept or reject it. If the task is admitted it joins the tail of one of the I queues. In addition, the DM controls the *fraction of effort* devoted by the *server* to the task at the head of queue i , for each i . Denote holding cost per time unit and rejection cost per customer of class- i as h_i and r_i , respectively. Denote reciprocal mean service time by μ_i . We refer to the two elements of control as *admission control* and *scheduling control*. The problem considered is to minimize the combination of holding and rejection costs. The motivation for this setting is inspired by the results demonstrated in [3]. However, there is crucial difference which is expressed in the buffer structure.

We assume a critical load condition and observe the model at the diffusion scale. In the scaling (diffusion) limit, the heavy traffic limits of Queuing control problem (QCP) turn to be a *Brownian control problem* (BCP). It is shown in [3] that there is an equivalence of an I -dimensional BCP and *reduced BCP* (RBCP), where a *workload* is a one-dimensional controlled state process. See references therein for additional discussion on BCP reduction.

The specific one-dimensional RBCP is related to *Hamilton-Jacobi-Bellman* (HJB) equation which, in our setting, takes the form of an ordinary differential equation. The solution to this problem was analyzed by Harrison and Taksar [10], as a singular control problem for a Brownian motion (BM), and is given by a reflected BM (RBM), and by a

free boundary point of the workload, denoted by \mathbf{x}^* . Thus, we are interested in the solution to the one-dimensional *workload control* in the interval of the form $[0, \mathbf{x}^*]$.

Hence, we treat Harrison-Taksar free boundary problem, specializing in the shared buffer. While the admission control is based on the simple indexing of $r_i\mu_i$ and the free boundary point \mathbf{x}^* , the scheduling control is complex. Namely, the difficulty comes from the shared buffer constraint, and is expressed in understanding the structure of the holding cost of the workload process. The expression for the workload holding cost, denoted by \bar{h} , (equation (27)), is motivated by the aforementioned equivalence of the BCP and RBCP. In particular, $\bar{h}(w)$, where w is a given workload point, is minimized. In contrary to the case of dedicated buffers, the solution for the holding cost in case of a shared buffer, is not immediate. Therefore, we formulate and solve this problem as a linear programming (LP) problem. Additionally, the solution is numerically demonstrated in this work.

Using the solution obtained from the LP and the simple $r\mu$ indexing, we construct a particular AO admission and scheduling policy which is specified by the free boundary point that is used in solving the BCP. To shortly summarize the policy, the domain of the queue lengths in a shared buffer can be geometrically represented by an I -simplex in the positive orthant. For example, in 2-dimensional system, it can be described by a triangle and in the 3-dimensional system by a tetrahedron. We show that the solution is such that the queue length process always follows one of the *edges* of the simplex. Hence, because of this structure of the queue length, we merely term it as *triangular* policy.

The indexes of h_i and μ_i are used for scheduling. We make a comparison between the system of dedicated buffers for all task classes, the case which analyzed in [3] (we name it as a *rectangular* case), and the shared buffer system which is analyzed here, the *triangular* case. We demonstrate that the priority indexing differs from the one used for the rectangular case and, in particular, from the $c\mu$ priority rule ([6]). In contrary, as we analyze in section 5, the resulting solution shows that there are *at most two classes that can be concurrently present in the shared buffer*. The key for setting the indexes stems from partitioning of the workload (one-dimensional) scale as follows. First, the buffer is filled up with tasks of the index, which is picked by the lowest $h_i\mu_i$. This is the first index. Next, then the buffer is full and the workload increases, tasks of second index class are accumulated, at expense of tasks of the first one. Once the workload continues to grow, the classes with higher indexes gradually replace those with the lower indexes. Thus, there are $J \leq I$ workload intervals, such that a unique pair of classes is designated to each interval. The indexes of the classes which fill the buffer are found by the LP.

Following the LP solution at the limit, we formulate the asymptotically optimal policy, such that when the workload level is below \mathbf{x}^* , all arrivals are admitted, with high probability. That is, forced rejections, which occur when the buffer size constraint is reached, have low probability. This is done by assigning dynamic priorities, such that classes which do not belong to the pair of classes which fill the buffer are immediately served. As a result, nearly all rejections occur when the workload exceeds \mathbf{x}^* , and only from one class, which is identified by the lowest $r\mu$ product. Under the AO policy, the I -dimensional queue length process converges to the process solving the RBCP. This convergence is a form of a *state space collapse* (SSC). Namely, the queue length process limits are dictated by the workload process limit. Note that the full statement of AO is accomplished by proving that the BCP value function is a lower bound on the limit inferior of QCP costs under any sequence of policies. This part is sufficiently general and is not covered in this paper. (See in [3], Theorem 1 for the details and the proof).

Various models associated with BCP were treated in [7], [11], [8], [9], [12] and others. Examples of characterization of the BCP and RBCP value functions as solutions to HJB equations are found in [1], [2]. Scheduling policies, such as $c\mu$ rule and extensions can be also studied from [6] and [18], and references therein. A general setting of SSC was considered in [5] and [19].

We finalize the introduction by bringing the practical motivation for the described setting. Our primary interest in this problem comes from recent developments in the application area of cloud computing. In particular, the constantly growing intensity of incoming, outgoing and traversing traffic in the public cloud (e.g. Amazon) makes the diffusion approximation, which we use as our main analytical tool, rather practical. We bring two concrete practical examples. First, consider a hybrid cloud, where a private cloud (namely, a *local server*) has a given capacity and memory limits. The tasks are served by sharing the computing effort among various task types. Once the incoming tasks cannot be queued for the local processing, they are rejected from a private cloud and sent to a public cloud, where a fixed charge per usage applies. In this context, [3] only treated the case where locally processed tasks compete for the server, having dedicated storage resources. In our case, the problem introduces additional motivation; namely, *tasks of different types compete with each other over the storage resources* as well. In this work, the common storage is modeled by a shared buffer. The second example refers to virtual machines (VM) allocation, a procedure which is performed by a hypervisor *residing on the side of the cloud's hardware*. The VM allocation is based on a common

computational resource, and their total number is naturally limited by hardware constraints. Since VM for various types of tasks are instantiated by a single hypervisor, one can view this resource as a buffer. Different types of tasks have different communication and computation demands which are expressed in differentiated service effort. (Assume the overhead and order of VM release and allocation are embodied in the service rate.) The rejection stands for inability to allocate an additional VM, while the rejection price stands for the VM migration cost. For further details on modeling of hybrid cloud applications, control and additional practical examples for the model analyzed in this work the reader is referred to [16], [15] and references therein.

Notation

We will use the following notation. Given $k \in \mathbb{N}$, $\{e^{(i)}, i = 1, \dots, k\}$ denote the standard basis in \mathbb{R}^k . For $x \in \mathbb{R}$, $x^+ = \max(x, 0)$. For $a, b \in \mathbb{R}^k$, $a = (a_i)_{i=1, \dots, k}$, $b = (b_i)_{i=1, \dots, k}$, we denote $\|a\| = \sum_{i=1}^k |a_i|$ and $a \cdot b = \sum_{i=1}^k a_i b_i$. For $y : \mathbb{R}_+ \rightarrow \mathbb{R}^k$ and $T > 0$, $\|y\|_T = \sup_{t \in [0, T]} \|y(t)\|$. The modulus of continuity of y is given by

$$\bar{w}_T(y; \theta) = \sup\{\|y(s) - y(t)\| : s, t \in [0, T], |s - t| \leq \theta\}, \quad \theta, T > 0.$$

Denote $A[s, t] = A(t) - A(s)$ for any process A .

The structure of the rest of the paper is as follows. In sections 2 and 3 we bring the definitions, which, in most parts, closely follow the definitions from [3]. We start with the queuing and diffusion models. We next define the Brownian control problem (BCP), and the reduced Brownian control problem (RBCP). Proposition 3.1 shows the relation between these two problems. These components are consistent with [3], and we bring them here because we will use them in the proof in section 5. We analyze HT solution for the shared buffer and bring a detailed numerical example in section 4. Section 5 constitutes the formulation of nearly optimal scheduling and admission policy. In Section 6 we state and prove Theorem 6.1, which, altogether with the lower bound stated in [3], implies the AO of the proposed policy. We conclude in Section 7.

2 Queueing and diffusion models

We start with definitions of queueing and diffusion models. Consider a sequence of systems, indexed by a superscript $n \in \mathbb{N}$. The system has a single server and single shared buffer, where tasks of all classes can occupy exactly one slot per task. The capacity of the buffer is limited. Customers that arrive at the system are judged by the DM to either be accepted or rejected. Those that are accepted are queued in the common queue. Within each class, service is provided in the order of arrival. Processor sharing is allowed, in the sense that the server is capable of serving up to I customers (of distinct classes) simultaneously, but the service cannot be shared within the customers of the same class. Denote an *allocation vector* of size I , representing the fractions of effort dedicated to the classes, and is chosen from the set

$$\mathcal{B} := \left\{ \beta \in \mathbb{R}_+^I : \sum_{i \in \mathcal{I}} \beta_i \leq 1 \right\},$$

where, $\mathcal{I} = \{1, 2, \dots, I\}$.

We assume a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with expectation w.r.t. \mathbb{P} denoted by \mathbb{E} . Arrivals occur according to independent renewal processes. The parameters $\lambda_i^n > 0$, $i \in \mathcal{I}$, $n \in \mathbb{N}$, satisfying $\lambda_i^n = n\lambda_i + \sqrt{n}\hat{\lambda}_i + o(\sqrt{n})$, with fixed positive λ_i and $\hat{\lambda}_i \in \mathbb{R}$, represent the *reciprocal mean inter-arrival times* of class- i tasks in the n -th system. The parameters $\mu_i^n > 0$, $i \in \mathcal{I}$, $n \in \mathbb{N}$, satisfying $\mu_i^n = n\mu_i + \sqrt{n}\hat{\mu}_i + o(\sqrt{n})$, with fixed positive μ_i and $\hat{\mu}_i \in \mathbb{R}$, represent the *reciprocal mean service times* of class- i tasks in the n -th system. Let $\{IA_i(l) : l \in \mathbb{N}\}_{i \in \mathcal{I}}$ be independent sequences of strictly positive i.i.d. random variables with mean $\mathbb{E}[IA_i(1)] = 1$, $i \in \mathcal{I}$ and squared coefficient of variation $\text{Var}(IA_i(1))/\mathbb{E}[IA_i(1)]^2 = C_{IA_i}^2 \in (0, \infty)$. The number of arrivals of class- i customers up to time t , for the n -th system, is given by

$$A_i^n(t) = A_i(\lambda_i^n t), \quad \text{where} \quad A_i(t) = \sup \left\{ l \geq 0 : \sum_{k=1}^l IA_i(k) \leq t \right\}, \quad t \geq 0. \quad (1)$$

Let independent sequences $\{ST_i(l) : l \in \mathbb{N}\}_{i \in \mathcal{I}}$ of strictly positive i.i.d. random variables be given, with mean $\mathbb{E}[ST_i(1)] = 1$ and squared coefficient of variation $\text{Var}(ST_i(1))/\mathbb{E}[ST_i(1)]^2 = C_{ST_i}^2 \in (0, \infty)$. The time required to complete the l -th service to a class- i customer in the n -th system is given by $ST_i(l)/\lambda_i^n$ units of time dedicated by

the server to this class. This is also known as *potential service time* processes, that is

$$S_i^n(t) = S_i(\mu_i^n t), \quad \text{where} \quad S_i(t) = \sup \left\{ l \geq 0 : \sum_{k=1}^l ST_i(k) \leq t \right\}, \quad t \geq 0. \quad (2)$$

$S_i^n(t)$ is the number of class- i jobs completed by the time when the server has dedicated t units of time to work on jobs of this class. We assume the sequences $\{IA_i\}$ and $\{ST_i\}$ are independent. We assume the *critical load condition*

$$\sum_{i \in \mathcal{I}} \rho_i = 1, \quad \text{where} \quad \rho_i = \frac{\lambda_i}{\mu_i}, \quad i \in \mathcal{I}. \quad (3)$$

The number of class- i rejections until time t and customers present at time t in the n -th system is denoted by $Z_i^n(t)$ and $X_i^n(t)$, correspondingly. Since rejections occur only at times of arrival, we have

$$Z_i^n(t) = \int_{[0,t]} z_s^{n,i} dA_i^n(s) \quad (4)$$

for some process $z^{n,i}$.

We call $X^n = (X_i^n)_{i \in \mathcal{I}}$ the queue length process. We assume, deterministic $X_i^n(0)$ and that no partial service has been provided to any of the tasks present in the system at time zero. Let $B^n = (B_i^n)_{i \in \mathcal{I}}$ be a process taking values in the set \mathcal{B} . Then

$$T_i^n(t) = \int_0^t B_i^n(s) ds \quad (5)$$

gives the time devoted to class- i customers up to time t . The number of service completions of class- i jobs during the time interval $[0, t]$ is given by

$$D_i^n(t) = S_i^n(T_i^n(t)). \quad (6)$$

We thus have

$$X_i^n(t) = X_i^n(0) + A_i^n(t) - D_i^n(t) - Z_i^n(t) = X_i^n(0) + A_i^n(t) - S_i^n(T_i^n(t)) - Z_i^n(t), \quad t \geq 0. \quad (7)$$

It is assumed that $B^n, S^n, Z^n, D^n, X^n, T^n$ have RCLL sample paths. Next define a rescaled version of the processes at diffusion scale as

$$\begin{aligned} \hat{A}_i^n(t) &= \frac{A_i^n(t) - \lambda_i^n t}{\sqrt{n}}, & \hat{S}_i^n(t) &= \frac{S_i^n(t) - \mu_i^n t}{\sqrt{n}}, & i \in \mathcal{I}, \\ \hat{Z}^n(t) &= \frac{Z^n(t)}{\sqrt{n}}, & \hat{X}^n(t) &= \frac{X^n(t)}{\sqrt{n}}. \end{aligned} \quad (8)$$

The shared buffer structure is specified as follows

$$\mathcal{X} = \{y \in \mathbb{R}_+^I : \sum_i y_i \leq b\}$$

We assume that the rescaled initial condition $\hat{X}^n(0)$ also lies in \mathcal{X} . The buffer constraint is always met, namely:

$$\hat{X}^n(t) \in \mathcal{X}, \quad t \geq 0, \text{ a.s.} \quad (9)$$

The rejection mechanism assures the condition above by rejecting arrivals occurring at a time t when $(X^n(t-) + e^{(i)})/\sqrt{n} \notin \mathcal{X}$. We refer to these rejections as *forced rejections*. In our setting, we distinguish them from admission/rejection decisions which are part of the control process. Note that the actual un-normalized buffer size scales like \sqrt{n} .

The control process $U^n = (Z^n, B^n)$, which is determined based on observations from the past (and present) events in the system, is defined as follows.

Definition 2.1: (Admissible control, QCP) Fix $n \in \mathbb{N}$ and consider fixed processes (A^n, S^n) given by (1) and (2). A process $U^n = (Z^n, B^n)$, taking values in $\mathbb{R}_+^I \times \mathcal{B}$, having RCLL sample paths with the processes $Z_i^n, i \in \mathcal{I}$ having nondecreasing sample paths and given in the form (4), is said to be an admissible control for the n -th system if the following holds. Let the processes T^n, D^n, X^n be defined by the A^n and S^n and control processes, (A^n, S^n) and (Z^n, B^n) , via equations (5), (6) and (7), respectively. Then

- (Z^n, B^n) is adapted to the filtration $\sigma\{A_i^n(s), D_i^n(s), i \in \mathcal{I}, s \leq t\}$;
- One has a.s., that, for all $i \in \mathcal{I}$ and $t \geq 0$,

$$X_i^n(t) = 0 \quad \text{implies} \quad B_i^n(t) = 0. \quad (10)$$

An admissible control under which the scaled version \hat{X}^n of X^n satisfies (9) is said to satisfy the buffer constraint.

We denote the class of all admissible controls $U^n = (Z^n, B^n)$ satisfying the buffer constraints, by \mathcal{U}^n . Fix $\alpha > 0$, $h \in (0, \infty)^I$ and $r \in (0, \infty)^I$. For each $n \in \mathbb{N}$ consider the cost

$$J^n(U^n) = \mathbb{E} \left[\int_0^\infty e^{-\alpha t} [h \cdot \hat{X}^n(t) dt + r \cdot d\hat{Z}^n(t)] \right] \quad (11)$$

The QCP value is given by

$$V^n = \inf_{U^n \in \mathcal{U}^n} J^n(U^n). \quad (12)$$

Denote by $\theta^n = (\theta_i^n)_{i \in \mathcal{I}}$, $\theta_i^n = 1/\mu_i^n$, and $\theta = (\theta_i)_{i \in \mathcal{I}}$, $\theta_i = 1/\mu_i$. The process $\theta^n \cdot X^n$, referred to as *workload*, its normalized version $\theta^n \cdot \hat{X}^n$ and its formal limit, $\theta \cdot X$, will take part in the state-space collapse.

3 The Brownian control problems

We address now the *limit problems*. First, we show that the scaled processes defined in the previous section give rise to the limit processes of the queue lengths. Hence, the definition of the I -dimensional BCP of the queue lengths follows. Next, we define the one-dimensional BCP of the workload. Finally, we show the *equivalence of the value functions* of these two problems.

Using (3), (7) and the definition of the rescaled processes, the following identity holds for $i \in \mathcal{I}$ and $t \geq 0$:

$$\hat{X}_i^n(t) = \hat{X}_i^n(0) + \hat{W}_i^n(t) + \hat{Y}_i^n(t) - \hat{Z}_i^n(t), \quad (13)$$

where, denoting $m_i = \hat{\lambda}_i - \rho_i \hat{\mu}_i$, $m_i^n = \frac{\lambda_i^n - \rho_i \mu_i^n}{\sqrt{n}} = m_i + o(1)$,

$$\hat{W}_i^n(t) = \hat{A}_i^n(t) - \hat{S}_i^n(T_i^n(t)) + m_i^n t, \quad (14)$$

and

$$\hat{Y}_i^n(t) = \frac{\mu_i^n}{\sqrt{n}} (\rho_i t - T_i^n(t)). \quad (15)$$

Since $\sum_i \rho_i = 1$ and one always has $\sum_i B_i^n(t) \leq 1$, it follows that

$$\theta^n \cdot \hat{Y}^n \quad \text{is a nonnegative, nondecreasing process.} \quad (16)$$

We are interested in limit problem, where the diffusion coefficient n is taken to infinity. As far as the real system is concerned, we view configuration where task arrival rate and, consequently, the service rate grow large. This connection makes the following model to be of a particular practical interest. Formally, the limits of (13)–(16) and (11)–(12) give rise to a control problem associated with diffusion. Consider equation (13). We assume that the scaled initial conditions $\hat{X}^n(0)$ converge to x as $n \rightarrow \infty$. Next, the centered, rescaled renewal process \hat{A}_i^n [resp., \hat{S}_i^n] converges weakly to a BM starting from zero, with zero mean and diffusion coefficient $\sqrt{\lambda_i} C_{IA^i}$ [resp., $\sqrt{\mu_i} C_{ST^i}$] (see Section 17 of [4]). Assume the processes involved in (13) give rise to a limiting BCP. Then it follows \hat{Y}^n in (15) are order one as $n \rightarrow \infty$ and, consequently, $T^n(t)$ converge to ρt . Thus, applying the time change in the second term of r.h.s. of (14), one sees that \hat{W}^n converges to (m, σ) -BM starting from zero, with drift vector $m = (m_i)_{i \in \mathcal{I}}$ and diffusion matrix $\sigma = \text{diag}(\sigma_i)$, where

$$\sigma_i^2 := \lambda_i C_{IA^i}^2 + \mu_i C_{ST^i}^2 \rho_i = \lambda_i (C_{IA^i}^2 + C_{ST^i}^2).$$

As for \hat{Y}^n , it gives rise to a process Y such that $\theta \cdot Y$ is nonnegative and nondecreasing. The process \hat{Z}^n gives rise to a process $\theta \cdot Z$ with nonnegative, nondecreasing components.

The BCP

Definition 3.2: (Admissible control, BCP) An admissible control for the initial condition $x_0 \in \mathcal{X}$ is a filtered probability space $(\Omega', \mathcal{F}', \{\mathcal{F}'_t\}, \mathbb{P}')$ for which there exists an (m, σ) -BM, W , and positive process $U = (Y, Z)$, with RCLL sample paths, such that the following conditions hold:

- W, Y and Z are adapted to $\{\mathcal{F}'_t\}$;
- For $0 \leq s < t$, the increment $W_t - W_s$ is independent of \mathcal{F}'_s under \mathbb{P}' ;

$$\theta \cdot Y \text{ and } Z_i, i = 1, \dots, I, \text{ are nondecreasing;} \quad (18)$$

- With
$$X(t) = x_0 + W(t) + Y(t) - Z(t), \quad t \geq 0, \quad (19)$$

one has

$$X(t) \in \mathcal{X} \quad \text{for all } t, \mathbb{P}'\text{-a.s.} \quad (20)$$

In what follows we will always assume that (Y, Z) are admissible controls. Denote the class of such controls as $\mathcal{A}(x_0)$, where x_0 stands for the initial condition. Let

$$J(x_0, Y, Z) = \mathbb{E} \left[\int_0^\infty e^{-\alpha t} [h \cdot X_t dt + r \cdot dZ(t)] \right]. \quad (21)$$

The BCP is to find (Y, Z) that minimize $J(Y, Z)$ and achieve the value

$$V(x_0) = \inf_{(Y, Z) \in \mathcal{A}(x_0)} J(x, Y, Z). \quad (22)$$

The RBCP

The reduced one-dimensional problem is obtained as follows. Multiply equation (19) and the processes involved in it by θ . Denote $\bar{x}_0 = \theta \cdot x_0$, $\bar{m} = \theta \cdot m$ and $\bar{\sigma}^2 = \sum \theta_i^2 \sigma_i^2$. Let

$$\mathbf{x} = \max\{\theta \cdot \xi : \xi \in \mathcal{X}\}. \quad (23)$$

Definition 3.3: (Admissible control, RBCP) An admissible control for the initial condition $\bar{x}_0 \in [0, \mathbf{x}]$ is a filtered probability space $(\Omega', \mathcal{F}', \{\mathcal{F}'_t\}, \mathbb{P}')$ for which there exist an $(\bar{m}, \bar{\sigma})$ -BM, \bar{W} , and a positive process $\bar{U} = (\bar{Y}, \bar{Z})$ with RCLL sample paths, such that the following conditions hold:

- \bar{W}, \bar{Y} and \bar{Z} are adapted to $\{\mathcal{F}'_t\}$;
- For $0 \leq s < t$, the increment $\bar{W}_t - \bar{W}_s$ is independent of \mathcal{F}'_s under \mathbb{P}' ;
- $$\bar{Y} \text{ and } \bar{Z} \text{ are nondecreasing;} \quad (24)$$

- With
$$\bar{X}(t) = \bar{x}_0 + \bar{W}(t) + \bar{Y}(t) - \bar{Z}(t), \quad t \geq 0, \quad (25)$$

one has

$$\bar{X}(t) \in [0, \mathbf{x}] \quad \text{for all } t, \mathbb{P}'\text{-a.s.} \quad (26)$$

We write $\bar{\mathcal{A}}(\bar{x}_0)$ for the class of admissible controls for the initial condition \bar{x}_0 . Given $(\bar{Y}, \bar{Z}) \in \bar{\mathcal{A}}(\bar{x}_0)$, let

$$\bar{J}(\bar{x}_0, \bar{Y}, \bar{Z}) = \mathbb{E} \left[\int_0^\infty e^{-\alpha t} [\bar{h}(\bar{X}_t) dt + \bar{r} d\bar{Z}(t)] \right], \quad (27)$$

where the holding cost and the rejection cost for the one-dimensional problem are defined as follows:

$$\bar{h}(w) = \min\{h \cdot \xi : \xi \in \mathcal{X}, \theta \cdot \xi = w\}, \quad w \in [0, \mathbf{x}],$$

$$\bar{r} = \min\{r \cdot z : z \in \mathbb{R}_+^I, \theta \cdot z = 1\}.$$

One interprets \bar{r} as a minimal rejection penalty per unit of work. Note that \bar{h} is convex by convexity of the set \mathcal{X} . Note also that as members of $(0, \infty)^I$, θ and h cannot be orthogonal, thus $\bar{h}(w) > 0$ for any $w > 0$. Since $\bar{h}(0) = 0$, it follows that \bar{h} is strictly increasing. Let

$$\bar{V}(\bar{x}_0) = \inf_{(\bar{Y}, \bar{Z}) \in \bar{\mathcal{A}}(\bar{x}_0)} \bar{J}(\bar{x}_0, \bar{Y}, \bar{Z}).$$

The following definitions will relate the two problems as follows. See, that the extremal points of the set $\{z \in \mathbb{R}_+^I : \theta \cdot z = 1\}$ are precisely $\theta_i^{-1} e^{(i)}$, namely $\mu_i e^{(i)}$, $i \in \mathcal{I}$. Hence there exists (at least one) i^* such that $\zeta = \mu_{i^*} e^{(i^*)}$ satisfies

$$\zeta \in \arg \min_z \{r \cdot z : z \in \mathbb{R}_+^I, \theta \cdot z = 1\}.$$

Fix such i^* and the corresponding ζ . See that i^* can be expressed via

$$r_{i^*} \mu_{i^*} = \min_i r_i \mu_i. \quad (28)$$

Next, let $\gamma : [0, \mathbf{x}] \rightarrow \mathcal{X}$ be Borel measurable, satisfying

$$\gamma(w) \in \arg \min_{\xi} \{h \cdot \xi : \xi \in \mathcal{X}, \theta \cdot \xi = w\}, \quad w \in [0, \mathbf{x}]. \quad (29)$$

By definition, $\gamma(w) \in \mathcal{X}$, $\theta \cdot \gamma(w) = w$, and $h \cdot \gamma(w) = \bar{h}(w) \leq h \cdot \xi$ for every $\xi \in \mathcal{X}$ for which $\theta \cdot \xi = w$. Hence, as it will become clear soon, these definitions of the costs will imply the equivalence of the value functions of BCP and RBCP.

Remark 3.1: One observes that the rejection process in the problem of dedicated buffers (denote it as a *rectangular case*) and that of a shared buffer (which we term as a *triangular case*) will follow the same rule. However, as opposed to [3], the solution for $\bar{h}(w)$ in the case of shared buffer is not immediately understood. We analyze this in the sequel of this section.

We now state the key proposition which determines the relation between the BCP and RBCP. This proposition, together with proposition 4.2 stated in sequel, leads to the formulation of the AO control policy.

Proposition 3.1: Let $x_0 \in \mathcal{X}$ and $\bar{x}_0 = \theta \cdot x_0$.

i. Given an admissible control $(\Omega', \mathcal{F}', \{\mathcal{F}'_t\}, \mathbb{P}', W, Y, Z)$ for x for the (multidimensional) BCP, define $(\bar{W}, \bar{X}, \bar{Y}, \bar{Z})$ by $(\theta \cdot W, \theta \cdot X, \theta \cdot Y, \theta \cdot Z)$. Then $(\bar{Y}, \bar{Z}) \in \bar{\mathcal{A}}(\bar{x}_0)$ and $\bar{J}(\bar{x}_0, \bar{Y}, \bar{Z}) \leq J(x_0, Y, Z)$.

ii. Conversely, let an admissible control $(\Omega', \mathcal{F}', \{\mathcal{F}'_t\}, \mathbb{P}', \bar{W}, \bar{Y}, \bar{Z})$ for \bar{x}_0 for the RBCP be given, and assume the probability space supports an (m, σ) -BM W . Assume W is $\{\mathcal{F}'_t\}$ -adapted and satisfies $\theta \cdot W = \bar{W}$ and (17). Construct (X, Y, Z) by

$$X(t) = \gamma(\bar{X}(t)), \quad Z(t) = \zeta \bar{Z}(t), \quad (30)$$

$$Y(t) = X(t) - x_0 - W(t) + Z(t). \quad (31)$$

Then $(Y, Z) \in \mathcal{A}(x_0)$, and $J(x_0, Y, Z) \leq \bar{J}(\bar{x}_0, \bar{Y}, \bar{Z})$.

iii. Consequently, $V(x_0) = \bar{V}(\bar{x}_0)$.

Note that the proposition holds for any convex \bar{h} . Hence, we skip the proof, which the interested reader can find in [3].

4 The Harrison-Taksar free boundary problem

We now focus on the one-dimensional problem. The function \bar{V} is $C^2[0, \mathbf{x}]$ and solves the following Bellman equation

$$\begin{cases} \left[\frac{1}{2} \bar{\sigma}^2 f'' + \bar{m} f' - \alpha f + \bar{h} \right] \wedge f' \wedge [\bar{r} - f'] = 0, & \text{in } (0, \mathbf{x}), \\ f'(0) = 0, & f'(\mathbf{x}) = \bar{r}. \end{cases} \quad (32)$$

This result has been demonstrated by Harrison and Taksar [10]. They showed that an optimal control is such that the process \bar{X} is a RBM. We introduce a *Skorohod map*, next. Let $a > 0$. The Skorohod map on the interval $[a, b]$, denoted by $\Gamma_{[a,b]}$, is map $D([0, \infty) : \mathbb{R}) \rightarrow D([0, \infty) : \mathbb{R})^3$. It solves *Skorohod Problem*, and its solution is a map $\psi \rightarrow (\varphi, \eta_1, \eta_2)$, for a given ψ , a triplet $(\varphi, \eta_1, \eta_2)$, such that

$$\varphi = \psi + \eta_1 - \eta_2, \quad \varphi(t) \in [a, b] \text{ for all } t, \quad (33)$$

$$\eta_i \text{ are nonnegative and nondecreasing, and } \int_{[0, \infty)} 1_{(a,b]}(\varphi) d\eta_1 = \int_{[0, \infty)} 1_{[a,b)}(\varphi) d\eta_2 = 0. \quad (34)$$

See [14] for existence and uniqueness of solutions, and continuity and further properties of the map. In particular, it is well-known that $\Gamma_{[a,b]}$ is continuous in the uniformly-on-compacts topology.

The following proposition is mostly a result of [10].

Proposition 4.2: The function \bar{V} is C^2 on $[0, \mathbf{x}]$ and solves (32) uniquely among all C^2 functions. Denote $\mathbf{x}^* = \inf\{y \in [0, \mathbf{x}] : \bar{V}'(z) = \bar{r} \text{ for } z \in [y, \mathbf{x}]\}$. Then $\mathbf{x}^* \in (0, \mathbf{x})$. Fix $\bar{x}_0 \in [0, \mathbf{x}]$. Let \bar{W} be an $(\bar{m}, \bar{\sigma})$ -BM and let \bar{X} , \bar{Y} and \bar{Z} be the corresponding RBM on $[0, \mathbf{x}^*]$ and boundary terms for 0 and \mathbf{x}^* , defined as

$$(\bar{X}, \bar{Y}, \bar{Z}) = \Gamma_{[0, \mathbf{x}^*]}(\bar{x}_0 + \bar{W}). \quad (35)$$

Then (\bar{Y}, \bar{Z}) is optimal for $\bar{V}(\bar{x}_0)$, i.e., $\bar{J}(\bar{x}_0, \bar{Y}, \bar{Z}) = \bar{V}(\bar{x}_0)$.

The reader is referred to [3] for the proof. An optimal control for the BCP stems from propositions 3.1 and 4.2.

4.1 Solution for \bar{h}

As mentioned, the specifics imposed by the shared property of a buffer implies non-trivial solution to the one-dimensional holding cost. The structure of the workload process is such that $\bar{X} = \theta \cdot X$ is given as a RBM on $[0, \mathbf{x}^*]$, where the free boundary point \mathbf{x}^* is dictated by the Bellman equation. The multidimensional queue length process X is recovered from \bar{X} by $X = \gamma(\bar{X})$.

Recall that the shared buffer domain has the form

$$\mathcal{X} = \{x \in \mathbb{R}_+^I : 0 \leq \sum_{i=1}^I x_i \leq b, i \in \mathcal{I}\}, \quad (36)$$

for some fixed $b > 0$. As mentioned, there is no difference between the cases of the shared and the dedicated models, in rejection structure. Namely, the BCP solution implies that in the scaled model, rejections should occur only when the scaled workload exceeds the level \mathbf{x}^* , and only from class i^* , for which $r_i \mu_i$ is minimal. Next, the relation

$$\hat{X}^n = \gamma(\theta \cdot \hat{X}^n) + o(1) \quad (37)$$

between the queue length and workload processes must hold. We solve for the minimizing curve γ , where \mathcal{X} takes the form (36). Equation (29) can be written as

$$\gamma(w) \in \arg \min_x \{h \cdot x : 0 \leq x_i, 0 \leq \sum_{i=1}^I x_i \leq b \text{ and } \theta \cdot x = w\}, \quad w \in [0, \mathbf{x}].$$

To compare a rectangular system and a triangular system, one sees that the description of \bar{h} in the latter case is not trivial. This is because the minimization is done under restriction of the shared buffer. Namely, in rectangular case, it is clear that as long as the workload grows, the buffers with lower priorities are filled up until their limits are reached. However, in triangular case, one have to explicitly solve for \bar{h} in order to understand which task types are being preferred for each workload point.

Hence, we find \bar{h} by the linear program (LP). Assume first the buffer is full, and the total workload is w . We later extend the solution for the workload points where the buffer is partially empty. The standard LP form is written

$$\begin{aligned} & \text{minimize } \sum_{i=1}^I h_i x_i \quad \text{w.r.t} \\ & \sum_{i=1}^I \theta_i x_i = w, \quad \sum_{i=1}^I x_i = b \end{aligned} \quad (38)$$

Without loss of generality, consider two classes, denoted by class 1 and class 2, such that $\theta_2 > \theta_1$, and $\theta_1 b < w$, $\theta_2 b > w$. These conditions are general, and will be needed for the convenience of the canonical representation of the LP. Rewrite (38) in matrix canonical representation:

$$\begin{aligned} & \text{minimize} \\ & \frac{h_1(w - \theta_2 b)}{\theta_1 - \theta_2} + \frac{h_2(\theta_1 b - w)}{\theta_1 - \theta_2} + (h_3 - h_1 \frac{\theta_3 - \theta_2}{\theta_1 - \theta_2} - h_2 \frac{\theta_1 - \theta_3}{\theta_1 - \theta_2})x_3 + \dots + (h_I - h_1 \frac{\theta_I - \theta_2}{\theta_1 - \theta_2} - h_1 \frac{\theta_1 - \theta_I}{\theta_1 - \theta_2})x_I \\ & \text{w.r.t} \\ & \begin{pmatrix} x_1 & \frac{(\theta_3 - \theta_2)}{\theta_1 - \theta_2} x_3 & \frac{(\theta_4 - \theta_2)}{\theta_1 - \theta_2} x_4 & \dots & \frac{(\theta_I - \theta_2)}{\theta_1 - \theta_2} x_I \\ & x_2 & \frac{(\theta_1 - \theta_3)}{\theta_1 - \theta_2} x_3 & \frac{(\theta_1 - \theta_4)}{\theta_1 - \theta_2} x_4 & \dots & \frac{(\theta_1 - \theta_I)}{\theta_1 - \theta_2} x_I \end{pmatrix} = \begin{pmatrix} \frac{b\theta_2 - w}{\theta_2 - \theta_1} \\ \frac{w - b\theta_1}{\theta_2 - \theta_1} \end{pmatrix} \end{pmatrix} \quad (39)$$

In the case all the coefficients of the new objective function in the above display are positive, by [Theorem 3.4.1, [17]] the minimum for \bar{h} is achieved at

$$x = \left\{ \frac{b\theta_2 - w}{\theta_2 - \theta_1}, \frac{w - b\theta_1}{\theta_2 - \theta_1}, 0, \dots, 0 \right\},$$

and is expressed by

$$\bar{h} = \frac{h_1(w - \theta_2 b)}{\theta_1 - \theta_2} + \frac{h_2(\theta_1 b - w)}{\theta_1 - \theta_2}$$

Alternatively, assume $(h_3 - h_1 \frac{\theta_3 - \theta_2}{\theta_1 - \theta_2} - h_2 \frac{\theta_1 - \theta_3}{\theta_1 - \theta_2})$ is negative. Then, we act according to the simplex algorithm, as presented in [17] and perform a pivot operation. This gives the following problem:

$$\begin{aligned} & \text{minimize} \\ & \frac{h_1(w - \theta_3 b)}{\theta_1 - \theta_3} + \frac{h_2(\theta_1 b - w)}{\theta_1 - \theta_3} + (h_3 - h_1 \frac{\theta_2 - \theta_3}{\theta_1 - \theta_3} - h_2 \frac{\theta_1 - \theta_2}{\theta_1 - \theta_3})x_2 + \dots + (h_I - h_1 \frac{\theta_I - \theta_3}{\theta_1 - \theta_3} - h_1 \frac{\theta_1 - \theta_I}{\theta_1 - \theta_3})x_I \\ & \text{w.r.t} \\ & \begin{pmatrix} x_1 & \frac{(\theta_2 - \theta_3)}{\theta_1 - \theta_3} x_2 & \frac{(\theta_4 - \theta_3)}{\theta_1 - \theta_3} x_4 & \dots & \frac{(\theta_I - \theta_3)}{\theta_1 - \theta_3} x_I \\ \frac{(\theta_1 - \theta_2)}{\theta_1 - \theta_3} x_2 & x_3 & \frac{(\theta_1 - \theta_4)}{\theta_1 - \theta_3} x_4 & \dots & \frac{(\theta_1 - \theta_I)}{\theta_1 - \theta_3} x_I \end{pmatrix} = \begin{pmatrix} \frac{b\theta_3 - w}{\theta_3 - \theta_1} \\ \frac{w - b\theta_1}{\theta_3 - \theta_1} \end{pmatrix} \end{pmatrix} \quad (40)$$

One sees that the solution to this problem has a similar form. Hence, in the case all coefficients in the function to be minimized are positive, the minimum for \bar{h} is achieved at this time

$$x = \left\{ \frac{b\theta_3 - w}{\theta_3 - \theta_1}, 0, \frac{w - b\theta_1}{\theta_3 - \theta_1}, 0, \dots, 0 \right\},$$

and is expressed by

$$\bar{h} = \frac{h_1(w - \theta_3 b)}{\theta_1 - \theta_3} + \frac{h_3(\theta_1 b - w)}{\theta_1 - \theta_3}$$

See from [17], (section 3.8) that there is a finite number of pivot steps which brings the canonical representation above to the form with objective function with non-negative coefficients. (There are additional options, such that the objective function is unbounded and/or degenerate cases which we rule out since they imply non-physical properties of the problem setting.) Therefore, for any w , the minimal $\bar{h}(w)$ is achieved when *only tasks belonging to at most two classes are present in the buffer*. Note that the solution above is not necessarily unique, as more than one pair of θ_i may satisfy the conditions above. However, once initial workload is at 0, the tasks of some class, denote it for a moment as j , are accumulated first. This class has the lowest h_j/θ_j . That is, the first type of task to be accumulated in the buffer follows the well-known $c\mu$ rule, till it holds $w = b\theta_j$. We derive next the rule how to find the two tasks which minimize the \bar{h} , for $w > b\theta_j$. See that the maximal workload associated with tasks of (the cheapest) type j , is the lowest workload possible provided the buffer is full. Now assume that additional amount of workload worth of ϵ is added. This workload addition is translated into reduction of number of tasks of type j and addition of tasks of other types. Denote the number of tasks subtracted from type j as X . Then, the number of added tasks of all other types is $\sum_{i=1, i \neq j}^I c_i X$, for non-negative c_i , such that $\sum_{i=1, i \neq j}^I c_i = 1$. Total workload change is given by

$$\epsilon = \sum_{i=1, i \neq j}^I c_i X \theta_i - X \theta_j$$

Then, the total number of tasks of type j that were displaced (by saying displaced we mean served, and no additional tasks of that type were admitted) is

$$X = \frac{\epsilon}{\sum_{i=1, i \neq j}^I c_i \theta_i - \theta_j}$$

We find the optimal cost associated with the workload addition worth of ϵ , denote it as P .

$$P = \sum_{i=1, i \neq j}^I c_i X h_i - X h_j = \frac{\epsilon (\sum_{i=1, i \neq j}^I c_i h_i - h_j)}{\sum_{i=1, i \neq j}^I c_i \theta_i - \theta_j}$$

Alternatively, cost addition "per workload worth of ϵ " is

$$P/\epsilon = \frac{\sum_{i=1, i \neq j}^I c_i h_i - h_j}{\sum_{i=1, i \neq j}^I c_i \theta_i - \theta_j} \quad (41)$$

The following lemma is a straight-forward consequence of the theorem [3.4.1] mentioned above:

Lemma 4.1: The optimal combination of c_i with respect to (41) is such that there exists one $c_k = 1, k \neq j$ and $c_i = 0$ for $i \neq k$.

See that once the buffer is full with tasks of some type j , only tasks of the type with the *lowest ratio* $\frac{h_k - h_j}{\theta_k - \theta_j}$ are admitted. That is, the additional cost accumulated due to the workload increase is minimal. Heuristically, as the workload continues to increase, this admission pattern continues till the buffer is full with tasks of type k , while all tasks of type j have been displaced. Then, the tasks of the next type which comply to the similar condition are accumulated, instead of those of type k . The addition of the workload cannot go further if the free boundary is reached. Note that the first class to be admitted (when the buffer is not full) is the one with the lowest ratio $\frac{h_k}{\theta_k}$. We name this heuristically described procedure as *order of accumulation*. Figure 1 demonstrates the description above, in 4 possible cases, for systems with 3 and 2 classes.

Remark 4.2: For more consistency, one can define $\theta_0 = 0$ and $h_0 = 0$. Then, the expression $\frac{h_k - h_j}{\theta_k - \theta_j}$ will hold when the buffer is not full as well.

We use these results to give a formal definition for the AO policy in the next section.

4.1.1 Numerical example of Harrison-Taksar problem and its solution.

In what follows we present the numerical example. The graphs below demonstrate the numerical solution of Harrison-Taksar problem for a shared buffer with maximal occupancy given by $b = 125$ tasks, and other parameters given in

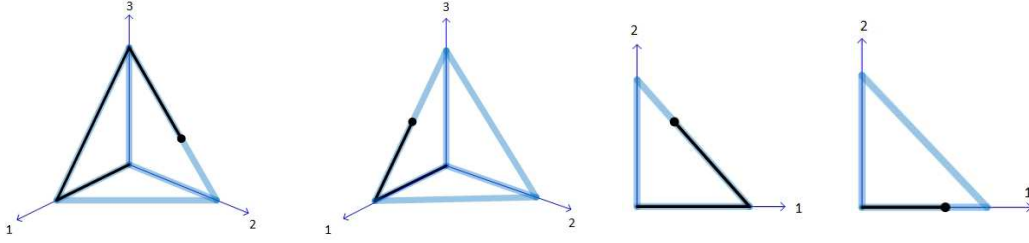


Fig. 1: The case of 2 sharing classes (right), and 3 sharing classes (left). Rejection point may be reached before the buffer with least priority is admitted. For a higher rejection level, the curve continues along the boundary.

Tab. 1: Table of parameters of 3 sharing tasks

Class	Holding cost	Rejection cost	μ	λ	$r\mu$
Class I	1390	962.5	1.80	0.60	1732.4
Class II	1050	700	2.20	0.73	1539.1
Class III	733	875	2.80	0.93	2450.3

Table 1. We assumed $\hat{\lambda}_i = \hat{\mu}_i = 0$ for all i (so that $\bar{m} = 0$), that $\bar{\sigma}^2 = 0.91$, and took the discount parameter $\alpha = 10$. The ordering of $r_i\mu_i$ is such that class 2 is the less expensive as far as rejections are concerned. Thus, this class is rejected at the free boundary.

The Bellman equation takes the form

$$\begin{cases} [2.19f'' - 10f + \bar{h}] \wedge f' \wedge [1539.19 - f'] = 0, & \text{in } (0, 69.45), \\ f'(0) = 0, & f'(69.45) = 1539.19. \end{cases} \quad (42)$$

The function \bar{h} is defined by

$$\bar{h}(w) = \min \left\{ \sum_{i=1}^3 h_i \xi_i : \xi \in \mathcal{X}, \sum_{i=1}^3 \theta_i \xi_i = w \right\}, \quad w \in [0, 69.45],$$

where $\mathcal{X} = [0, 125]$ and $\theta_i = \mu_i^{-1}$, which by numerical solution of the corresponding LP is translated into the following

$$\bar{h}(w) \approx \begin{cases} \frac{732.9 \cdot w}{0.36} & 0 \leq w \leq 44.64, \\ \frac{732.9(125 \cdot 0.45 - w)}{0.45 - 0.36} + \frac{1050(w - 125 \cdot 0.36)}{0.45 - 0.36} & 44.64 < w \leq 56.85, \\ \frac{1050(125 \cdot 0.56 - w)}{0.56 - 0.45} + \frac{1390(w - 125 \cdot 0.45)}{0.56 - 0.45} & 56.85 < w \leq 68.06. \end{cases}$$

Table 2 demonstrates the solution to \bar{h} and the selection of classes for each of the workload intervals. This selection takes part in setting the γ . Note that we used remark 4.2.

j	$\frac{h_1-h_j}{\theta_1-\theta_j}$	$\frac{h_2-h_j}{\theta_2-\theta_j}$	$\frac{h_3-h_j}{\theta_3-\theta_j}$
0	2501.8	2308.7	2052.4
3	3310.3	3245.6	
2	3.3730		

Tab. 2: Order of task accumulation. j stands for the class which, in case the workload grows, fills up the buffer. The optimal selection, within each row, is in bold. As long as the buffer is not full, only class I is admitted. Next, tasks of class II are admitted. As the workload grows, tasks of class I are displaced, while tasks of class II fill up the buffer. Finally tasks of class III are accumulated, till free boundary is reached. We put blanks where the classes were already chosen for lower workload intervals.

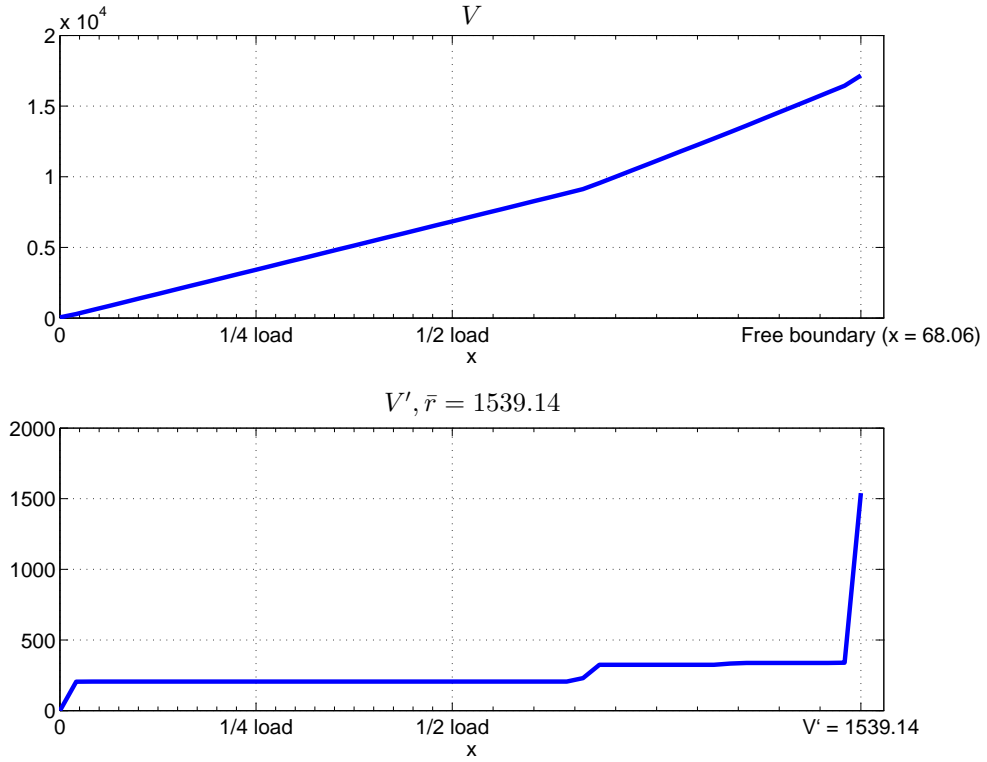


Fig. 2: V and V' within the free boundary. See that the maximal value of V' is equal to \bar{r} at free boundary.

The optimal curve γ is given by the following

$$\gamma(w) \approx \begin{cases} \frac{w}{44.64} 125 [0, 0, 1] & 0 \leq w \leq 44.64, \\ \frac{125 * 0.45 - w}{0.45 - 0.36} [0, 0, 1] + \frac{w - 125 * 0.36}{0.45 - 0.36} [0, 1, 0] & 44.64 < w \leq 56.85, \\ \frac{125 * 0.56 - w}{0.56 - 0.45} [1, 0, 0] + \frac{w - 125 * 0.45}{0.56 - 0.45} [0, 1, 0] & 56.85 < w \leq 68.06. \end{cases}$$

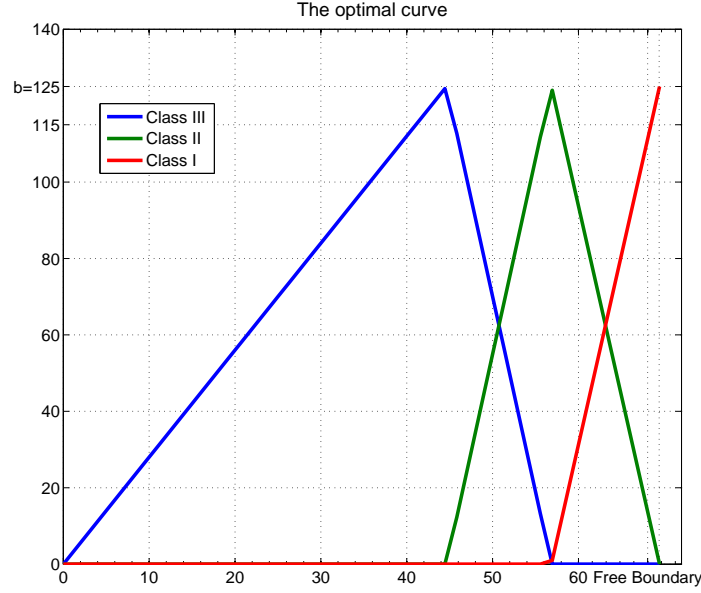


Fig. 3: Classes which form the optimal curve γ . X-axis stands for the workload, while Y-axis shows the number of tasks of each class present in the buffer at this workload. The optimal queue-lengths at free boundary are $\{0, 27.5595, 97.4405\}$ (dashed vertical line).

5 A nearly optimal policy

We introduce the curve γ for the triangular domain, given by (36). The parameter \mathbf{x} associated with the RBCP is given by $\theta_m b$, where $m = \arg \max_i \theta_i$. This corresponds to the buffer being full with most processor-consuming tasks. The tasks of this class contribute the maximal per-task workload. In what follows, we label the classes according to their *order of accumulation*, which was heuristically described in previous section. Again, assume that the workload grows from 0 to \mathbf{x} . This ordering will form the *accumulation priorities* $p(j)$, and it will be used to formulate the mapping from w to $\gamma(w)$. The numbering $j \in \{1, \dots, I\}$, refers to the workload points \hat{w}_j , where the buffer is solely full with tasks of some type $i = p(j)$, such that $\hat{w}_j < \hat{w}_{j+1}$. (Note that in sequel we redefine this ordering in order to obtain $i = p(i)$.) $\hat{w}_0 = 0$, stands for the zero workload. \hat{w}_1 is defined as w_1 in the previous subsection, i.e the minimal load then the buffer is full. Namely,

$$\hat{w}_1 = \theta_n b, \quad n = \arg \min_i h_i \mu_i, \quad i \in \{1, \dots, I\}$$

The calculation of $p(j) > 1$ is performed recursively.

$$p(j) = \arg \min_i \frac{h_i - h_{p(j-1)}}{\theta_i - \theta_{p(j-1)}} : h_i > h_{p(j-1)}, \theta_i > \theta_{p(j-1)}$$

Observe that the restriction $h_i > h_{p(j-1)}, \theta_i > \theta_{p(j-1)}$ means that no class can be chosen twice. (For otherwise that i would be chosen for the lower $p(j)$). In addition, the condition on θ_i assures that the workload added by each task of class $p(j)$ is higher than that of task of class $p(j-1)$. See that $|p(j)| \leq I$. Note that some classes can never be optimal, (for example, those with comparatively high h_i and low θ_i) and thus be never accumulated. However, the tasks of class with maximal θ_i are always assigned to the last j . Note that by remark 4.2 one can define $h_{p(0)} = 0$ and $\theta_{p(0)} = 0$ to have the order of accumulation started with an empty buffer.

Define two groups of tasks:

$$\begin{aligned}\mathcal{E} &= \{i | \exists j : p(j) = i\} \\ \mathcal{D} &= \{i | \nexists j : p(j) = i\}\end{aligned}$$

Denote $p_m = |\mathcal{E}|$. The group \mathcal{D} constitutes the task types which are always at low accumulation priority. See that \mathcal{D} can be an empty set. Next, for simplicity of notation, we *reorder* the class indexing such that $i = p(i)$, if $i \in \mathcal{E}$ and $i = \{p_m + 1, p_m + 2, \dots, I\}$, if $i \in \mathcal{D}$. That is, the tasks are rearranged in the order of their accumulation in the buffer (i.e. the order of i is equivalent to the order of j before the reordering). Denote $J = |\mathcal{E}|$. The workload intervals are denoted by ω_j .

$$\omega_j = [\hat{w}_{j-1}, \hat{w}_j), \quad j > 1 \quad (43)$$

Given $w \in [0, \mathbf{x}]$, $(j, \xi_l, \xi_h) = (j, \xi_l, \xi_h)(w)$ are determined by $w \in \omega_j$ and we use the solution to (39) to set

$$\xi_h(w) = \frac{w - b\theta_{j-1}}{\theta_j - \theta_{j-1}}, \quad \xi_l(w) = b - \xi_h(w), \quad (44)$$

in case $j > 1$, and

$$\xi_l(w) = 0, \quad \xi_h(w) = w/\theta_1, \quad (45)$$

in case $j = 1$. With this notation, γ is given by

$$\gamma(w) = \xi_h\theta_j + \xi_l\theta_{j-1}.$$

Clearly, the curve, which is expressed by γ defined above, can lie along the boundary $\partial\mathcal{X}$ of \mathcal{X} , where it holds $\partial^+\mathcal{X} := \{x \in \mathcal{X} : \sum_i x_i = b\}$, i.e. the buffer is full. Note that γ is the solution to the BCP which, in addition, comes in concert with the free boundary solution found from the RBCP, according to which only rejections at the workload level \mathbf{x}^* are allowed. Thus, these two properties of the solution can be seen as contradicting, when treating the QCP. Hence, we propose a policy which approximates γ by introducing an alternative curve, which is closed to γ and is bounded away from the buffer limit boundary at the same time. Let $\varepsilon \in (0, b)$ be given. Let $a = b - \varepsilon$, and $a^* := \mathbf{x}^* \wedge (\theta_J a) < \mathbf{x} = \theta_J b$. In the case ε is small, we have then $a^* = \mathbf{x}^*$ (unless $\mathbf{x}^* = \mathbf{x}$). Note that the buffer can be full for various values of ξ_l, ξ_h . Therefore, we define additional margins that can vary according to the workload w . Let $\varepsilon_l = \varepsilon_l(w)$ and $\varepsilon_h = \varepsilon_h(w)$. For each w we set ε_l and ε_h as follows

$$\begin{aligned}\varepsilon_l &= \begin{cases} \varepsilon_l = \min\{\varepsilon/2, \xi_l\}, & \text{if } \xi_l < \xi_h, \\ \varepsilon_l = \varepsilon/2, & \text{if } \xi_l \geq \xi_h, \xi_h > \varepsilon/2 \\ \varepsilon_l = \varepsilon - \xi_h, & \text{if } \xi_l \geq \xi_h, \xi_h \leq \varepsilon/2 \end{cases} \\ \varepsilon_h &= \varepsilon - \varepsilon_l\end{aligned} \quad (46)$$

Let $\chi_l = \xi_l - \varepsilon_l$ and $\chi_h = \xi_h - \varepsilon_h$. The approximation $\gamma^a : [0, \mathbf{x}] \rightarrow \mathcal{X}$ of γ is as follows. For $w \in [0, \theta_J a)$, the variables $j = j(w)$ and $\chi_l = \chi_l(w), \chi_h = \chi_h(w)$ are determined via

$$w = \theta_{j-1}\chi_l + \theta_j\chi_h, \quad j \in \{1, 2, \dots, J\}, \quad \chi_l + \chi_h \in [0, a), \quad (47)$$

and

$$\gamma^a(w) = \chi_l e^{(j-1)} + \chi_h e^{(j)}. \quad (48)$$

Note that the triplet (j, χ_l, χ_h) is unique. We will refer to it as the *representation* (j, χ_l, χ_h) of w via (47). Denote $w_\chi = \theta_l\chi_l + \theta_h\chi_h$ and $w_\xi = \theta_l\xi_l + \theta_h\xi_h$, where θ_l and θ_h refer to the classes which are associated with ξ_l and ξ_h , correspondingly. We need the function γ^a to be continuous on $[w_\chi, w_\xi]$ and satisfy the relation $\theta \cdot \gamma^a(w) = w$. Hence, we define the linear interpolation between the points w_χ and w_ξ as follows:

$$\gamma^a(w) = a + \frac{w - w_\chi}{w_\xi - w_\chi}(b - a), \quad w \in [w_\xi, w_\chi]. \quad (49)$$

We next specify the policy. $Z^n(t)$ defines the rejection policy, while $B^n(t)$ describes the server allocation policy, as a function of $X^n(t)$. In what follows the policy is set considering the index correction after the reordering above.

Rejection policy:

The multidimensional rejection process Z has only one nonzero component, namely the i^* -th component, which increases only when $\bar{X} \geq \mathbf{x}^*$. This structure is translated to the queueing model. Hence, we use it to construct the rejection component of the asymptotically optimal policy. The latter occurs when $\theta \cdot \hat{X}^n \geq a^*$. The forced rejections only occur in order to keep the buffer size constraint.

Service policy:

For each $x \in \mathcal{X}$ define the low priority:

$$\mathcal{L}(x) = \{j | x_j < \chi_h, \ j - 1 | x_{j-1} < \chi_l\}$$

That is, the tasks of type j and $j - 1$, where $j = j(w)$ have always low priority as long as their quantity is below χ_l and χ_h , correspondingly. The complement set defines the high priority classes:

$$\mathcal{H}(x) := \mathcal{I} \setminus \{\mathcal{L}(x)\}.$$

Denote $\mathcal{H}^+(x) = \{i \in \mathcal{H}(x) : x_i > 0\}$ and $\mathcal{L}^+(x) = \{i \in \mathcal{L}(x) : x_i > 0\}$. The policy is to allocate service to all classes within $\mathcal{H}^+(x)$ equal to a fraction proportional to the corresponding traffic intensities. Classes within $\mathcal{L}(x)$ receive no service, with exception of the moments then $\mathcal{H}^+(x)$ is empty.

$$\rho_i^L(x) = \begin{cases} 0, & \text{if } 1_{(x_i=0) \cup (\mathcal{H}^+(x) \neq \emptyset)} = 1, \\ \frac{\rho_i 1_{\{i \in \mathcal{L}^+(x)\}}}{\sum_{k \in \mathcal{L}^+(x)} \rho_k}, & \text{if } \mathcal{H}^+(x) = \emptyset \end{cases}, \quad \rho_i^H(x) = \begin{cases} 0, & \text{if } x_i = 0, \\ \frac{\rho_i 1_{\{i \in \mathcal{H}^+(x)\}}}{\sum_{k \in \mathcal{H}^+(x)} \rho_k}, & \text{if } \mathcal{H}^+(x) \neq \emptyset, \end{cases} \quad (50)$$

Define $\rho'_i(x) = \rho_i^L(x) \cdot 1_{i \in \mathcal{L}} + \rho_i^H(x) \cdot 1_{i \in \mathcal{H}}$. Then for each t ,

$$B^n(t) = \rho'(\hat{X}^n(t)). \quad (51)$$

Note that when $\mathcal{H}^+(x) \neq \emptyset$,

$$\rho'_i(x) > \rho_i \quad \text{for all } i \in \mathcal{H}^+(x). \quad (52)$$

That is, all classes which receive service are allocated a fraction of effort strictly greater than their traffic intensity. See that $\sum_i B_i^n = 1$ whenever \hat{X}^n is nonzero. Hence, the proposed policy is work conserving. Figure 4 demonstrates two cases of service policy for 6 classes. The reader may also consider to go back to Table 2 to review the example of the order of accumulation.

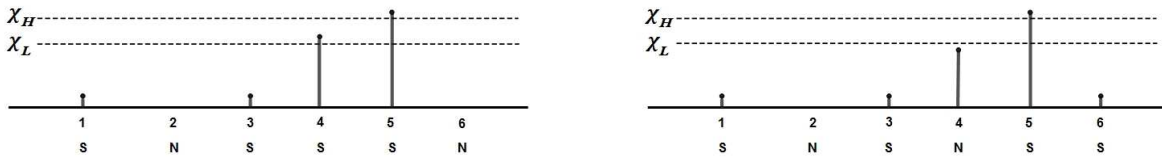


Fig. 4: Schematic example of service allocation. The figures depict possible states $\hat{X}^n(t) = x$. Classes denoted by S are being served, classes denoted by N are not being served.

Remark 5.3: The tasks of type $i > J$ (if exist) have always high priority once arrive to the system. This is because these tasks are never optimal to have them accumulated in the buffer, according to the solution of \bar{h} . However, one of these task types may be chosen to be rejected at free boundary, according to the \bar{r} .

Remark 5.4: Note, that the rejection policy resembles that of the rectangular case, because the calculation of \bar{r} follows the same rule. The class i^* is the only class which is subject to the non-forced rejection.

6 State space collapse.

In this section, we prove the main theorem which states the upper bound for the value function under proposed policy. The theorem is written as follows.

Theorem 6.1: For each $\varepsilon > 0$ and n , denote the policy constructed above by $U^n(\varepsilon)$.

Then,

$$\limsup_{n \rightarrow \infty} J^n(U^n(\varepsilon)) \leq V(x_0) + \alpha(\varepsilon), \text{ where } \alpha(\varepsilon) \rightarrow 0 \text{ as } \varepsilon \rightarrow 0.$$

Remark 6.5: The combination of Theorem 6.1 above and Theorem 3.1 in [3] (which states the general lower bound) provides the AO of the policy $U^n(\varepsilon)$.

Proof.

For fixed ε in $U^n(\varepsilon)$, write $U^n = (Z^n, B^n)$ and denote by τ^n the time of the first forced rejection. We perform most of the analysis on the processes before τ^n is reached. We show that the cost of QCP, under proposed policy, weakly converges to the value function associated with BCP solution as $n \rightarrow \infty$. We divide the proof into two major steps. First step shows that the workload process $\theta^n \cdot \hat{X}^n$ converges to RBM. This results in that only rejections from class i^* occur, and only when $\theta^n \cdot \hat{X}^n \approx a^*$. The second step, shows that \hat{X}^n lies close to the minimizing curve at all times. Hence, the running cost is locally minimized. This establishes that in any finite time, τ^n is not reached.

We prove for the case where the system starts with initial condition close to the minimizing curve, namely

$$\hat{X}^n(0) - \gamma^a(\theta \cdot \hat{X}^n(0)) \rightarrow 0 \text{ as } n \rightarrow \infty, \text{ and } \theta^n \cdot \hat{X}^n(0) \in [0, a^*] \text{ for all } n \text{ large.} \quad (53)$$

The assumption on initial condition can be relaxed in a straightforward manner. Shortly, given a general initial condition, there will be a jump towards a position located on the curve. We skip the technical details which can be found in [3].

Denote some ε' , a distance from the minimizing curve $\gamma^a(\theta \cdot x)$, small enough to assure that forced rejections within that distance do not occur. Observe that $\sigma^n \leq \tau^n$, where

$$\sigma^n = \hat{\zeta}^n \wedge \zeta^n, \\ \hat{\zeta}^n = \inf\{t : X^{\#,n} \geq a^* + \varepsilon'\}, \quad \zeta^n = \inf\{t : \max_{i \leq I} |\Delta_i^n(t)| \geq \varepsilon'\}.$$

In this sense, $\hat{\zeta}^n$ refers to the time of violating of the free boundary of the workload, while ζ^n refers to the time when distance of number of tasks of any tasks-type from the optimality curve overcomes ε' .

Now multiply equation (13) by the vector $\theta^n = (1/\mu_i^n)_{i \in \mathcal{I}}$ and denote

$$W^{\#,n} = \theta^n \cdot \hat{W}^n, \quad X^{\#,n} = \theta^n \cdot \hat{X}^n, \quad Y^{\#,n} = \theta^n \cdot \hat{Y}^n, \quad Z^{\#,n} = \theta^n \cdot \hat{Z}^n. \quad (54)$$

We have

$$X^{\#,n} = X^{\#,n}(0) + W^{\#,n} + Y^{\#,n} - Z^{\#,n}. \quad (55)$$

Let $W^{\circ,n} := W^{\#,n}(\cdot \wedge \tau^n)$ denote the process $W^{\#,n}$ when stopped at the time τ^n . Define similarly $X^{\circ,n}$, $Y^{\circ,n}$ and $Z^{\circ,n}$.

The following lemma states that the free boundary is not violated (that is, $\hat{\zeta}^n$ is not reached).

Lemma 6.2: any subsequential limit $(\tilde{W}, \tilde{X}, \tilde{Y}, \tilde{Z})$ of the sequence $(W^{\circ,n}, X^{\circ,n}, Y^{\circ,n}, Z^{\circ,n})$ is C -tight, and satisfies a.s.,

$$(\tilde{X}, \tilde{Y}, \tilde{Z}) = \Gamma_{[0, a^*]}[\tilde{x}_0 + \tilde{W}]. \quad (56)$$

The detailed proof is given in [3]. (See Theorem 4.1, *Step 1*).

Hence, we straightly proceed to prove our main result, which constitutes the proof of the state-space collapse. Our objective is to show that the multidimensional process \hat{X}^n lies close to the minimizing curve. That is, as $n \rightarrow \infty$,

$$\Delta^n(t) := \hat{X}^n(t) - \gamma^a(X^{\#,n}(t)) \Rightarrow 0, \quad (57)$$

uniformly on compacts.

It suffices to show that $\mathbb{P}(\sigma^n < T) \rightarrow 0$, for any small $\varepsilon' > 0$ and any T . Fix ε' and T . Thanks to the fact that $\sigma^n \leq \tau^n$,

$$\mathbb{P}(\sigma^n < T) \leq \mathbb{P}(\hat{\zeta}^n \wedge \zeta^n \leq T \wedge \tau^n) \leq \mathbb{P}(\hat{\zeta}^n \leq T \wedge \tau^n) + \mathbb{P}(\zeta^n \leq T \wedge \tau^n). \quad (58)$$

From lemma 6.2 it follows that $\mathbb{P}(\hat{\zeta}^n \leq T \wedge \tau^n) \rightarrow 0$ as $n \rightarrow \infty$. It therefore suffices to prove the following lemma.

Lemma 6.3: $\mathbb{P}(\zeta^n \leq T \wedge \tau^n) \rightarrow 0$ as $n \rightarrow \infty$.

Proof.

On $\zeta^n \leq T \wedge \tau^n$ let $x^n := X^{\#,n}(\zeta^n) = X^{\circ,n}(\zeta^n)$ and let $j = j^n$ and χ_l^n, χ_h^n be the corresponding components from the representation (j, χ_l, χ_h) of x^n (with $w = x^n$).

The proof strategy is to define a covering of the workload domain $[0, \mathbf{x}]$ by intervals. Then, we assume w being present in one of these intervals and prove the lemma for all possible cases. More precisely, we show that it is sufficient to distinguish between four different types of intervals in order to treat all possibilities.

To that end, fix a positive integer $K = K(\varepsilon') = \lceil c_0/\varepsilon' \rceil$, where c_0 is a constant depending only on θ . (we treat this value at a later stage of the proof.) Define $K - 1$ intervals $\Xi_k = \mathbf{B}(k\varepsilon_1, \varepsilon_1)$, $k = 1, 2, \dots, K - 1$, where $\mathbf{B}(x, d)$ denotes a closed interval $[x - d, x + d]$ and $\varepsilon_1 = \mathbf{x}/K$. Let $\tilde{\Xi}_k = \mathbf{B}(k\varepsilon_1, 2\varepsilon_1)$.

We use the characterization of C -tightness as in Proposition VI.3.26 of [13] and apply it to $X^{\circ,n}$. Given $\delta > 0$ there exists $\delta' = \delta'(\delta, T, \varepsilon_1) > 0$, such that for all sufficiently large n ,

$$|X^{\circ,n}(s) - X^{\circ,n}(t)| \leq \varepsilon_1 \text{ for all } s, t \in [0, T], |s - t| \leq \delta', \text{ with probability at least } 1 - \delta. \quad (59)$$

Fix such δ and δ' . Denote by \mathbf{T}^n the interval $[(\zeta^n - \delta' \vee 0), \zeta^n]$. Define the following event,

$$\Omega^{n,k} = \{\zeta^n \leq T \wedge \tau^n, x^n \in \Xi_k, X^{\#,n}(t) \in \tilde{\Xi}_k \text{ for all } t \in \mathbf{T}^n\}. \quad (60)$$

By simple probabilistic manipulations, using that $X^{\#,n} = X^{\circ,n}$ on $[0, \tau^n]$, it can be shown that for all large n ,

$$\mathbb{P}(\zeta^n \leq T \wedge \tau^n) \leq \delta + \sum_k \mathbb{P}(\Omega^{n,k}), \quad (61)$$

We fix the index k , which is associated with the k -th interval in the workload covering, and analyze $\Omega^{n,k}$. We aim to show that $\mathbb{P}(\Omega^{n,k}) \rightarrow 0$ as $n \rightarrow \infty$, for each k . The value assigned by the policy to B^n (see (51)) remains fixed as \hat{X}^n varies within any of the intervals (ω_j) defined in (43). However, this is not necessary the case. There are several cases, which we separately define and treat as it follows below.

(I)

$\tilde{\Xi}_k \subset (0, a^*)$ and for all j , $\hat{w}_j \notin \tilde{\Xi}_k$. That is, \hat{X}^n remains in the same interval during the time window \mathbf{T}^n . Hence, the region ω_j is constant. To prove in this case, see that all points x in $\tilde{\Xi}_k$ are translated to the same j in the representation (j, χ_l, χ_h) of x , as in (47). Note that $j = j(k)$ depends on the interval k only, and does not vary with n . Also, $j = j^n$ under $\Omega^{n,k}$. We split the proof into two separate cases corresponding to the two groups of classes - $i \notin \{j-1, j\}$, $i \in \{j-1, j\}$. We also separately treat the case where $j = 1$, i.e. the buffer is not full. We start with the first group. Fix $i \notin \{j-1, j\}$. We estimate the probability that, on $\Omega^{n,k}$, $\zeta^n \leq T \wedge \tau^n$ occurs by having $\Delta^n(\zeta^n) \geq \varepsilon'$. More precisely, note that $\gamma_i^a(x^n) = 0$ (this follows from the policy definition in (50), i.e. because $i \notin \{j-1, j\}$). Then we will show that

$$\text{for every } \varepsilon'' \in (0, \varepsilon'), \quad \mathbb{P}(\Omega^{n,k} \cap \{\hat{X}_i^n(\zeta^n) > \varepsilon''\}) \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (62)$$

Note that γ^a is continuous and that $\Delta^n(0) \rightarrow 0$ as $n \rightarrow \infty$, by (53). Using the fact that the jumps of \hat{X}^n are of size $n^{-1/2}$, on the event indicated in (62) there must exist $\eta^n \in [0, \zeta^n]$ with the properties that

$$\hat{X}_i^n(\eta^n) < \varepsilon''/2, \quad \hat{X}_i^n(t) > 0 \text{ for all } t \in [\eta^n, \zeta^n]. \quad (63)$$

Define $\hat{\eta}^n = \eta^n \vee (\zeta^n - \delta')$. We examine now $\hat{X}_i^n[\zeta^n] - \hat{X}_i^n[\hat{\eta}^n] = \hat{X}_i^n[\hat{\eta}^n, \zeta^n]$, aiming to bound it using the modulus of continuity. Since the probability of the latter to be positive should go to zero, the probability that the difference

$\hat{X}_i^n[\hat{\eta}^n, \zeta^n]$ is positive will go to zero as well. See that due to the position of $\hat{X}^n(t)$ in interval $\tilde{\Xi}_k \subset (0, a^*)$ (by definition of case (I)) there are no rejection at free boundary at all times in time interval $[\hat{\eta}^n, \zeta^n]$. Hence, $\hat{Z}_i^n[\hat{\eta}^n, \zeta^n] = 0$. Since the buffer is full, we must show that the probability of the forced rejections goes to zero as $n \rightarrow \infty$ for all i . Assume that jobs of classes $j-1$ and j are bounded away from ξ_l and ξ_h respectively. Hence, the arrivals of class i are admitted and $Z_i^n[\hat{\eta}^n, \zeta^n] = 0$, with high probability. Observe that on the event in (63), during the time interval $[\eta^n, \zeta^n]$, i is always a member of $\mathcal{H}(\hat{X}^n)$. Recall (from definition in (5)) $T_i^n(t) = \int_0^t B_i^n(s)ds$, while by (51)–(52), $B_i^n(t) = \rho'_i(\hat{X}^n(t)) > \rho_i + c$, for some constant $c > 0$. Substitute in (15) and make time derivative

$$\frac{d}{dt} \hat{Y}_i^n(t) \leq \frac{d}{dt} \left(\frac{\mu_i^n}{\sqrt{n}} (\rho_i t - \int_0^t (\rho_i + c) ds) \right) = -\frac{\mu_i^n}{\sqrt{n}} c \quad (64)$$

Note that definition of $\hat{\eta}^n$ assures that the equations which follow next will be valid on $\Omega^{n,k}$, i.e. within the time interval \mathbf{T}^n . That is, by (61) and (65) we have that the time interval $[\hat{\eta}^n, \zeta^n]$ is bounded by δ' , while $\hat{X}_i^n[\hat{\eta}^n, \zeta^n]$ on it is bounded by ε_1 . Using these facts in (13) and substituting $(\zeta^n - \hat{\eta}^n)$ for the time interval in (64), we have

$$\hat{X}_i^n[\hat{\eta}^n, \zeta^n] = \hat{W}_i^n[\hat{\eta}^n, \zeta^n] - c \frac{\mu_i^n}{\sqrt{n}} (\zeta^n - \hat{\eta}^n). \quad (65)$$

Fix r_n , a positive sequence, such that $r_n \rightarrow 0$ and $r_n \sqrt{n} \rightarrow \infty$. We use this sequence to indicate the growth of $\zeta^n - \hat{\eta}^n$ with n , referring to two different cases. The C -tightness of \hat{W}_i^n provides a bound by means of modulus of continuity; that is, $w_T(\hat{W}_i^n; r_n) \geq \hat{W}_i^n[\eta^n, \zeta^n]$.

In the first case, we assume $\zeta^n - \eta^n < r_n$ and n is sufficiently large, such that $\hat{\eta}^n = \eta^n$. Moreover, $c \frac{\mu_i^n}{\sqrt{n}} (\zeta^n - \hat{\eta}^n)$ is finite and positive. Thus by definition of η^n , and by the assumption that jump of size $\varepsilon''/2$ happened, we have $\hat{X}_i^n[\hat{\eta}^n, \zeta^n] \geq \varepsilon''/2$. Consequently,

$$w_T(\hat{W}_i^n; r_n) \geq \hat{W}_i^n[\eta^n, \zeta^n] \geq \varepsilon''/2$$

must hold. However, the probability of this event goes to zero as $n \rightarrow \infty$.

In the second case we assume that $\zeta^n - \eta^n \geq r_n$. Hence, by (65) we omit $\varepsilon''/2$ and change sides,

$$2\|\hat{W}_i^n\|_T \geq \hat{W}_i^n[\hat{\eta}^n, \zeta^n] \geq c \frac{\mu_i^n}{\sqrt{n}} r_n \geq cr_n \sqrt{n},$$

for some constant $c > 0$. Observe that since $r_n > 0$, the probability of the event above goes to zero as well. Summarizing, the probability in (62) is bounded by

$$\mathbb{P}(w_T(\hat{W}_i^n; r_n) \geq \varepsilon''/2) + \mathbb{P}(2\|\hat{W}_i^n\|_T \geq cr_n \sqrt{n}), \quad (66)$$

which converges to zero as $n \rightarrow \infty$, by C -tightness of \hat{W}^n . This proves (62). Consequently, the classes which do not belong to the $\{j-1, j\}$ stay at zero with high probability. Moreover, forced rejections of these classes happen with low probability.

Note that in the case $i \notin \{j-1, j\}$ we demonstrated that $X^n(t)$ lies near zero. Hence, we could assume that $\gamma_j^a(X^{\circ,n}) = 0$. In order to prove for the case $j-1$ and j we aim to show,

$$\text{for every } \varepsilon'' \in (0, \varepsilon'), \quad \mathbb{P}(\Omega^{n,k} \cap \{\Delta_j^n(\zeta^n) > \varepsilon''\}) \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (67)$$

That is, classes $j-1$ and j are bounded away from ξ_l and ξ_h by predefined constants ε_l and ε_h , respectively. Therefore, we will show that the distance from the curve, $\Delta_j^n[\hat{\eta}^n, \zeta^n]$, goes to zero with high probability. For simplicity of notation, define $C^n(t) = \gamma_j^a(X^{\circ,n}(t))$. Then $\Delta_j^n = \hat{X}_j^n - C^n$. Recall that by policy definition for $i = j$,

$$\begin{aligned} j &\in \mathcal{H}(\hat{X}^n(t)) \text{ whenever } \hat{X}_j^n(\eta^n) \geq \chi_h, \\ j &\in \mathcal{L}(\hat{X}^n(t)) \text{ whenever } \hat{X}_j^n(\eta^n) < \chi_h. \end{aligned} \quad (68)$$

We refer to the first option first. Similarly to (63), there exists $\eta^n \leq \zeta^n$ such that

$$\hat{X}_j^n(\eta^n) < \chi_h + \varepsilon''/2, \quad \hat{X}_j^n(t) > \chi_h \text{ for all } t \in [\eta^n, \zeta^n]. \quad (69)$$

Again, we distinguish between two cases: $\zeta^n - \eta^n < r_n$ and $\zeta^n - \eta^n \geq r_n$. We have now two terms of modulus of continuity, referring to the processes W_i^n and C^n . Recall that we by definition, γ_j^a is continuous ((48)–(49)) and by

Lemma 6.2 $X^{\circ,n}(t)$ is C -tight. Hence, C^n is C -tight. Therefore, proving similarly as for $i \notin \{j-1, j\}$, the probability in (67) is bounded by

$$\mathbb{P}(w_T(\hat{W}_i^n; r_n) + w_T(C^n; r_n) \geq \varepsilon''/2) + \mathbb{P}(2\|\hat{W}_i^n\|_T + 2\|C^n\|_T \geq cr_n\sqrt{n}), \quad (70)$$

Note that in the second option of (68) class i is not served and thus jumps immediately to the point χ_h with high probability, as $n \rightarrow \infty$. This accomplishes the proof for $i = j$. The proof for $i = j-1$ is similar, with the only difference that χ_h is substituted by χ_l . We now separately treat the case $j = 1$. For this case $j-1 = 0$ and $\chi_l = 0$, which is treated as for the task classes $i \notin \{j-1, j\}$, and (62) holds. The proof for the j itself is similar to the case where $j > 1$ and (67) holds. Thus,

$$\text{for every } \varepsilon'' \in (0, \varepsilon'), \quad \mathbb{P}(\Omega^{n,k} \cap \{\Delta_{j-1}^n(\zeta^n) > \varepsilon''\}) \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (71)$$

We can now show that $\mathbb{P}(\Omega^{n,k}) \rightarrow 0$ as $n \rightarrow \infty$. See that by definition of γ , the minimizing cost, and by the convergence established for the workload process $\theta \cdot \gamma^a(\theta \cdot x) = \theta \cdot x$ for all $x \in \mathcal{X}$ and $\theta^n \rightarrow \theta$. Combining this with (62) (67), (71), we bound $|\Delta_i^n(\zeta^n)|$ as follows

$$\mathbb{P}(\Omega^{n,k} \cap \{\max_{i \leq I} |\Delta_i^n(\zeta^n)| > \varepsilon''\}) \rightarrow 0, \quad (72)$$

Since ε'' is arbitrarily small, it follows from the definition of ζ^n that $\mathbb{P}(\Omega^{n,k}) \rightarrow 0$ as $n \rightarrow \infty$.

(II)

We consider the case $\tilde{\Xi}_k \subset (0, a^*)$ but $\hat{w}_j \in \tilde{\Xi}_k$ for some $j \in \{1, 2, \dots, J\}$. That is, the buffer is *mostly* filled with tasks of class j . Let $(j^n(t), \xi^n(t))$ denote the representation (47) for $X^{\#,n}(t)$. In the case $j^n > 2$, in the time window \mathbf{T}^n , $j^n, (j-1)^n$ vary in the range of $j-1, j$ and $j+1$. More precisely, we have to analyze the fluctuation between $\{j-1, j\}$ and $\{j, j+1\}$. The case $j^n = 2$, stands for fluctuation between 1, (then buffer is not nearly full) and $\{1, 2\}$. In case $j^n = 1$ we have $\hat{w}_0 \in \tilde{\Xi}_k$ which is separately treated in case **(III)**.

We first treat the case where $2 \leq j < J$. In this case, the buffer is mostly filled with tasks of class j . See by (44) that if $w > \hat{w}_j$, the tasks of class j are associated with χ_l . Define

$$\chi_h^{(1)}(w) = \frac{w - a\theta_j}{\theta_{j+1} - \theta_j}, \quad \chi_l^{(1)}(w) = a - \chi_h^{(1)}(w),$$

Otherwise, if $w \leq \hat{w}_j$, the number tasks of type j are associated with χ_h . Define

$$\chi_h^{(2)}(w) = \frac{w - a\theta_{j-1}}{\theta_j - \theta_{j-1}}, \quad \chi_l^{(2)}(w) = a - \chi_h^{(2)}(w),$$

Observe, that in the first case the buffer is full with tasks of type $j-1, j$, while in the second case the buffer is full with tasks of type $j, j+1$. The way we treat this is by bounding Δ^n from above by a quantity that depends on ε_1 , rather than by an arbitrarily small ε'' . Recall that ε_1 refers to the bound applied for the C -tightness of $X^{\circ,n}$ in (59) and is used to set the size of each interval in the covering Ξ_k and $\tilde{\Xi}_k$.

Define $c_1 = 4/\tilde{\theta}_{\min}$ and $\tilde{\theta}_{\min} = \min_i(\theta_{i+1} - \theta_i)$, where for convenience, we denote $\theta_0 = 0$. We have for any $w \in \tilde{\Xi}_k$, $|w - \hat{w}_j| \leq 4\varepsilon_1$, since \hat{w}_j is also in $\tilde{\Xi}_k$. Now, if $w \leq \hat{w}_j$, then for coordinate j of the minimizing curve it holds $\gamma_j^a(w) = \chi_h^{(2)}$. In this case,

$$w = (a - \chi_l^{(2)})\theta_h + \chi_l^{(2)}\theta_l = \hat{w}_j - \chi_l^{(2)}(\theta_h - \theta_l)$$

We have $\chi_l^{(2)} \leq 4\varepsilon_1/(\theta_j - \theta_{j-1})$, and consequently $\gamma_{j-1}^a(w) \leq 4\varepsilon_1/(\theta_j - \theta_{j-1})$. Thus, $\gamma_j^a(w) = a - \chi_l^{(2)} > a - \frac{4\varepsilon_1}{\theta_j - \theta_{j-1}}$. This shows that on $\Omega^{n,k}$,

$$\gamma_j^a(X^{\#,n}(t)) \geq a - c_1\varepsilon_1, \quad t \in \mathbf{T}^n, \quad (73)$$

Otherwise, if $w \geq \hat{w}_j$, it follows that $\gamma_{j-1}^a(w) = \chi_l^{(1)}$, and

$$w = \chi_h^{(1)}\theta_h + (a - \chi_h^{(1)})\theta_l = \hat{w}_j + \chi_h^{(1)}(\theta_h - \theta_l)$$

We have $\chi_h^{(1)} \leq 4\varepsilon_1/(\theta_{j+1} - \theta_j)$ and consequently $\gamma_j^a(w) \leq 4\varepsilon_1/(\theta_j - \theta_{j-1})$. Thus, $\gamma_j^a(w) > a - \frac{4\varepsilon_1}{\theta_{j+1} - \theta_j}$. This shows that in this case on $\Omega^{n,k}$,

$$\gamma_{j-1}^a(X^{\#,n}(t)) \geq a - c_1\varepsilon_1, \quad t \in \mathbf{T}^n, \quad (74)$$

Consequently, by position of $X^{\#,n}$ associated with the region $\tilde{\Xi}_k$ in case (II), we have by (73) and (74) that the minimizing curve is bounded away from $X^{\#,n}$ by order of ε_1 . Now, (62) is valid for all $i \notin \{j-1, j, j+1\}$, by the proof given in case (I).

For the case $j^n = 2$, if $w \geq \hat{w}_1$, the proof is identical. In the case $w \leq \hat{w}_1$, by (45), $\gamma_1^a(w) = w/\theta_1$. Since $\hat{w}_1 = a\theta_1$, it follows $a\theta_1 - w \leq 4\varepsilon_1$ and $\gamma_2^a(w) \geq a - 4\varepsilon_1/\theta_1 \geq a - c_1\varepsilon_1$. This shows that (73) holds in this case as well. Finally, for the case $j^n = 1$, if $w > \hat{w}_j$ then the proof is similar to the cases $1 \leq j^n < J$. (the fact that γ_1^a may assume the value zero does not affect the proof). Note that (62) is valid for all $i > 2$, by the proof given in case (I). For the particular case $j^n = J$ we have $\hat{w}_J = a^*$, because this is the maximal buffer load. Thus, we have only the case where $w \leq \hat{w}_J$, which is treated in the similar way.

Combining all the estimates for all small ε'' ,

$$\mathbb{P}(\Omega^{n,k} \cap \{\max_{i \neq j} \Delta_i^n(\zeta^n) > \varepsilon''\}) \rightarrow 0.$$

The estimate (67) and the bounds (73),(74) give $\mathbb{P}(\Omega^{n,k} \cap \{\Delta_j^n(\zeta^n) > 2c_1\varepsilon_1\}) \rightarrow 0$ as $n \rightarrow \infty$. This gives

$$\mathbb{P}(\Omega^{n,k} \cap \{\max_{i \leq J} |\Delta_i^n(\zeta^n)| > 4c_1\varepsilon_1\}) \rightarrow 0,$$

as $n \rightarrow \infty$.

We now determine the constant c_0 used to define K . The objective is to determine the relation between ε' , which used to set ζ^n , the first time the difference between the scaled process and the minimizing curve is greater than ε' , and ε_1 . We do this in such a way that $4c_1\varepsilon_1 < \varepsilon'/2$. In particular, any constant $c_0 > 8c_1\mathbf{x} = 32\mathbf{x}/\tilde{\theta}_{\min}$ will do.

We have bounded $\Delta_i^n(\zeta^n)$ by arbitrarily small ε' in case (I) and by ε_1 in case (II), and set the relation between these two bounds. This way we obtain $\mathbb{P}(\Omega^{n,k}) \rightarrow 0$ as $n \rightarrow \infty$.

(III)

$0 \in \tilde{\Xi}_k$. The is no difference of this case from case (I) with exception that \hat{X}^n may be at zero at some t . The analysis in case (I) results in the same conclusion, namely $\mathbb{P}(\Omega^{n,k}) \rightarrow 0$ as $n \rightarrow \infty$.

(IV)

$a^* \in \tilde{\Xi}_k$. This case corresponds to the rejections at free boundary. It is treated by adding a negative term in the equations written in case (I). (Note that we can pick ε to be sufficiently small to avoid treating (II) here).

Having shown that $\mathbb{P}(\Omega^{n,k}) \rightarrow 0$ in all cases, using (61) and the fact that $\delta > 0$ is arbitrary completes the proof of the lemma.

As a consequence of the lemma and (58), we have $\mathbb{P}(\sigma^n < T) \rightarrow 0$ as $n \rightarrow \infty$. Since ε' is arbitrary, (57) is established.

To accomplish the proof one has to show the weak convergence of the costs, i.e.

$$\int_0^\infty e^{-\alpha t} [h \cdot \hat{X}^n(t) + \alpha r \cdot \hat{Z}^n(t)] dt \Rightarrow \int_0^\infty e^{-\alpha t} [h \cdot \gamma^a(\bar{X}^a(t)) + \alpha \bar{r} \bar{Z}^a(t)] dt,$$

and to show the uniform integrability of rejection and holding cost components to obtain

$$\lim_n J^n(U^n) = \lim_n \mathbb{E} \left[\int_0^\infty e^{-\alpha t} [h \cdot \hat{X}^n(t) + \alpha r \cdot \hat{Z}^n(t)] dt \right] = \mathbb{E} \left[\int_0^\infty e^{-\alpha t} [h \cdot \gamma^a(\bar{X}^a(t)) + \bar{r} \bar{Z}^a(t)] dt \right] = V(x_0; \varepsilon)$$

Then, the final result follows once $V(x_0, \varepsilon) \rightarrow V(x_0)$ as $\varepsilon \rightarrow 0$ is seen. These technical steps are covered in [3] and are omitted here. Therefore, this accomplishes the proof of the theorem.

7 Conclusion

In this paper, we analyzed the problem of service scheduling and admission of multi-class tasks arriving to the shared buffer. The main motivation for this problem comes from the growing utilization of cloud computing resources. We demonstrated that in the limit, the queuing control is expressed via one-dimensional reduced workload control problem. Using the fact the the value function of the reduced and multi-task problem in the limit are coincide, we formulated the policy for which we proved, that once applied to the multi-task scaled problem, it attains the upper bound for the value function. Hence, we proved that the AO policy constitutes a form of state-space collapse. In particular, we demonstrated that the shared buffer is either occupied by two classes of tasks, or it is not full. The classes of the tasks are determined by the workload.

The workload process notion has a particular importance in computing systems. Namely, it indicates the remained computing effort till the buffer becomes empty. The AO policy we presented has an advantage of being simple to implement, which is important in cloud computing, then the managing overhead is of sensible weight. Moreover, as the rate of incoming tasks and cloud computing capabilities increase, the policy becomes more close to the optimal one and the optimality gap decreases. We conclude that diffusion approximation method has applicative meaning of constantly growing importance in this context. We suggest that other cloud computing configurations and computer networks in general can be similarly treated.

References

- [1] R. Atar and A. Budhiraja. Singular control with state constraints on unbounded domain. *Ann. Probab.*, 34(5):1864–1909, 2006.
- [2] R. Atar, A. Budhiraja, and R. J. Williams. HJB equations for certain singularly controlled diffusions. *Ann. Appl. Probab.*, 17(5-6):1745–1776, 2007.
- [3] R. Atar and M. Shifrin. An asymptotic optimality result for the multiclass queue with finite buffers in heavy traffic. *ArXiv e-prints*, Dec. 2014.
- [4] P. Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999. ISBN 0-471-19745-9. x+277 pp. A Wiley-Interscience Publication.
- [5] M. Bramson. State space collapse with application to heavy traffic limits for multiclass queueing networks. *Queueing Systems Theory Appl.*, 30(1-2):89–148, 1998.
- [6] C. Buyukkoc, P. Varaiya, and J. Walrand. The $c\mu$ rule revisited. *Adv. in Appl. Probab.*, 17(1):237–238, 1985.
- [7] J. M. Harrison. Brownian models of queueing networks with heterogeneous customer populations. In *Stochastic differential systems, stochastic control theory and applications (Minneapolis, Minn., 1986)*, volume 10 of *IMA Vol. Math. Appl.*, pages 147–186. Springer, New York, 1988.
- [8] J. M. Harrison. Brownian models of open processing networks: canonical representation of workload. *Ann. Appl. Probab.*, 10(1):75–103, 2000.
- [9] J. M. Harrison. A broader view of Brownian networks. *Ann. Appl. Probab.*, 13(3):1119–1150, 2003.
- [10] J. M. Harrison and M. I. Taksar. Instantaneous control of Brownian motion. *Math. Oper. Res.*, 8(3):439–453, 1983.
- [11] J. M. Harrison and J. A. Van Mieghem. Dynamic control of Brownian networks: state space collapse and equivalent workload formulations. *Ann. Appl. Probab.*, 7(3):747–771, 1997.
- [12] J. M. Harrison and R. J. Williams. Workload reduction of a generalized Brownian network. *Ann. Appl. Probab.*, 15(4):2255–2295, 2005.
- [13] J. Jacod and A. N. Shiryaev. *Limit theorems for stochastic processes*, volume 288 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1987. ISBN 3-540-17882-1. xviii+601 pp.
- [14] L. Kruk, J. Lehoczy, K. Ramanan, and S. Shreve. An explicit formula for the Skorokhod map on $[0, a]$. *Ann. Probab.*, 35(5):1740–1768, 2007.

-
- [15] M. Shifrin. *Admission and Scheduling Control in Cloud Computing -Markov Decision Processes and Diffusion Approximations*. PhD thesis, Technion, 2013.
 - [16] M. Shifrin, R. Atar, and I. Cidon. Optimal scheduling in the hybrid-cloud. In *IFIP/IEEE International symposium on integrated network management*. IEEE, 2013.
 - [17] P. R. Thie and G. E. Keough. *An introduction to linear programming and game theory*. John Wiley & Sons.
 - [18] J. A. van Mieghem. Dynamic scheduling with convex delay costs: the generalized $c\mu$ rule. *Ann. Appl. Probab.*, 5(3):809–833, 1995.
 - [19] R. J. Williams. Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing Systems Theory Appl.*, 30(1-2):27–88, 1998.